NEW YORK UNIVERSITY STERN
Leonard N. Stern School of Business

INTRO TO DATA SCIENCE
DS-GA-1001

SYLLABUS - Fall 2015

Professor	**Brian d'Alessandro,** Information, Operations & Management Sciences Department
Office; Hours	TBD; By Appointment
Email	TBD
Telephone	NA
Classroom	TBD
Class time	TBD
First : Last Class	TBD
Final Quiz	TBD
Course Assistants CA Office Hours	TBD

1. Course Overview

Businesses, governments, and individuals create massive collections of data as a by-product of their activity. Increasingly, decision-makers and systems rely on intelligent technology to analyze data systematically in order to improve decision-making. In many cases automating analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data analysis technologies can be used to improve decision-making. We will study the fundamental principles and techniques of data mining, and we will examine real-world examples and cases to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work "hands-on" with the Python programming language and its associated data analysis libraries.

After taking this course you should:

1. ***Understand what a Data Scientist is.*** The roles of a professional Data Scientist come in many flavors. With this class you will be able to identify where you fit in the data science spectrum.

2. ***Approach applicable problems data-analytically.*** Think carefully & systematically about whether & how data can improve decision making across a wide set of applications.

3. ***Learn and practice the core tools of Data Science.*** Data Science is a discipline requiring tools beyond just statistics and computer program. This course will teach you the foundations of machine learning, model building and Python programming while putting them to use solving real-world problems.

2. Focus and interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and appropriate technical details of machine learning, data mining and data science. The emphasis primarily is on understanding the fundamental concepts and applications of data science. We will cover several algorithms though this is not an algorithms course, nor a course in machine learning or computational theory. Our aim rather is to present fundamental algorithms within the context of a larger data mining and decision making process.

I will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures/discussions on the fundamental material, discussions of business applications of the ideas and techniques, case discussions, student exercises, and demos.

You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including (unless otherwise directed) having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class. In general, we will follow Stern default policies unless I state otherwise. I will assume that you have read them and agree to abide by them:

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

The NYU Classes site for this course will contain lecture notes, reading materials, assignments, and late-breaking news. You should check the site daily, and I will assume that you have read all announcements and class discussion.

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, come to office hours to see me or the TAs, or ask on the discussion board. The discussion board is the preferred method of asking questions, as others may benefit from the answers being available on NYU Classes. As a corollary to this, please try to answer your classmates' questions. In grading your class participation I will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

Worth repetition: It is your responsibility to check NYU Classes (and your email) at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent.

3. Lecture Notes, Readings and Quizzes

This is a graduate course so we'll assume that you have self-motivation and discipline to keep up with the readings on your own merit. Nonetheless, we'll do the occasional quiz to test your understanding and absorption of the material.

Primary Books: The primary textbook for the class will be:

Data Science for Business: Fundamental principles of data mining and data analytic thinking Provost & Fawcett (O'Reilly, 2013).

The book is now available, and you can purchase it in the bookstore or online (see http://data-science-for-biz.com/).

This book covers the fundamental material that will provide the basis for you to think and communicate about data science and business analytics. We will complement the book with discussions of applications, relevant white papers and readings from the following additional books.

We will supplement the above book with selected chapters from the following.

The Elements of Statistical Learning
Hastie, Tibshirani, Friedman (Springer 2009)

Book
http://statweb.stanford.edu/~tibs/ElemStatLearn/
Free PDF
http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

Additionally, we have crafted a series of iPython notebooks to supplement the above books on various technical details (regarding Python and Math).

http://nbviewer.ipython.org/github/briandalessandro/DataScienceCourse/tree/master/ipython/

Additional Books:

The following books are not mandatory but are worth owning and mastering as part of your development as a data scientist. These particularly focus on doing data science with the Python language.

Python for Data Analysis: Data Wrangling with Pandas, Numpy and iPython
Wes McKinney (O'Reilly 2012)

http://shop.oreilly.com/product/0636920023784.do

Data Science from Scratch: 1st Principles with Python
Joel Grus, (O'Reilly 2015)

http://shop.oreilly.com/product/0636920033400.do

Readings: A lot of valuable reference material is not published as a text book but instead comes in the form of academic and conference white papers. Throughout this course, we'll review several papers as mandatory readings. We'll also post several more as recommended supplementary material. In general, learning to read through these types of papers will only help you in a career that requires continuous learning.

Lecture notes: In an effort to conserve resources, lecture materials will be handed out digitally. I expect you to ask questions about any material in the notes that is unclear after our class discussion and reading the book. Having the book frees up class time for more discussion of applications, cases, etc.—so many of your questions may be answered in the book. If not, please let me know! Depending on the direction our class discussion takes, we may not cover all material in the class notes for any particular session. If the notes and the book are not adequate to explain a topic we skip, you should ask about it on the discussion board. I will be happy to follow up.

Please don't hesitate to come and talk to me about what supplemental material might be best for you, if you want to go further on any topics covered in the course.

4. Requirements and Grading

The grade breakdown is as follows:

1. Homeworks: 30%
2. Term Project: 30%
3. Participation, Class Contribution and Quizzes: 15%
4. Final Quiz: 25%

At NYU Stern and the Center for Data Science we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving As and Bs, and only one or two perform only adequately or below and receive C's or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

Homework Assignments

The homework assignments are listed (by assignment date) in the class schedule below. Each homework comprises questions to be answered and/or hands-on tasks. Except as explicitly noted otherwise (see next paragraph), you are expected to complete your assignments on your own—without interacting with on the completion of your assignment. You are free of course to discuss the concepts with your classmates, and to discuss similar problems to the ones in the homework.

For the hands-on parts of the assignments (with Python), I encourage you to work with your group members and other classmates to understand how to get Python to do what you need to do, and then to complete your assignment on your own. So, for example, you could have a classmate help you do something similar, such that then you would be able to complete the assignment.

I hope with the support of me, the TAs, and your classmates, we operate under a "diligent attempt but limited frustration" policy: (1) If you get stuck on something, spend some time Googling to try to find the answer. If you seem to be moving forward, keep going. That search and discovery will pay off, both in terms of the direct learning about how to do what you need to do, and also in terms of your learning *how to find* such things out. (E.g., if you don't know what Stackoverflow is, you will learn!). BUT, (2) limit frustration—start your assignments early enough that if you run into a wall, you can just stop searching and ask about it. Let's say, if you feel like you have not moved forward after 15 minutes of being stuck, just stop and ask: your classmates, on the discussion board, to the TAs. If you don't get a solution, escalate it to me.

Completed assignments must be handed on blackboard at least one hour prior to the start of class on the due date (that is, by 5pm), unless otherwise indicated. Assignments will be graded and returned promptly. Answers to homework questions should be well thought out and communicated precisely, as if reporting to your boss, client, potential funding source. Avoid sloppy language, poor diagrams, irrelevant discussion, and irrelevant program output.

The hands-on tasks in the homework will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class. You will use these data to build and evaluate predictive models.

Pre-Requisites

This is a technical/programming-oriented version of the popular course Data Mining for Business Analytics. You will receive credit only for one of the two.

You need not be a hacker, must have some proficiency in programming to take this class. It will be sufficient to have taken one of Stern's prior Python-oriented classes (Dealing with Data, Practical Data Science, Programming in Python). Alternatively, students may have developed programming proficiency elsewhere—and can get permission from the instructor.

Hands on Programming

Data Science is not possible without some sort of programming knowledge. This class will involve hands-on assignments and demonstrations of executing data mining techniques with the Python programming language and its associated libraries. We will cover the installation of the iPython Notebook (http://ipython.org/notebook.html).

For those students with little to no programming experience (Python or not), it is highly advised to install iPython and start learning the language and platform in advance of the class. There are many online tutorials for getting started with the language.

IMPORTANT: ***You must have access to a computer on which you can install software. If you do not have such a computer, please see me immediately so we can make alternative arrangements.*** During class we will have live demos of certain analysis techniques in action, using the iPython programming environment.

Generally the Course Assistants should be the first point of contact for questions about and issues with the homeworks. The primary course assistant (see first page) will have the responsibility to make sure that all questions are answered in a timely fashion, but please make use of both, as we have staggered the office hours to provide broader coverage. *If they cannot help you to your satisfaction, please do not hesitate to come see me.*

Late Assignments
As stated above, assignments are to be submitted on NYU Classes at least *one hour prior* to the start of the class on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time.

Term Project
A term project report will be prepared by student teams. We will give you the instructions on how to form your teams. Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. You will submit various milestone deliverables through the course. We will discuss the project requirements in class.

Final Quiz
The final quiz will be a take-home to be completed during the days following the last class. The subject matter covered and the exact dates will be discussed in class.

Participation/Contribution/Attendance/Punctuality
Please see Section 2.

Regrading
If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., pages in the book, a copy of class notes, etc.). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go “in your favor” and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

***FOR STUDENTS WITH DISABILITIES*:** If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for

Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

Please read the policies for Stern courses

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

Please keep in mind the Stern Honor Code

http://www.stern.nyu.edu/mba/studact/mjc/hc.html

Class Schedule - Fall 2015

Class #	Date	Topics *(subject to change as class progresses)* **Most classes will also include a case study/guest/lab/demo/etc.**	Required Readings	Deliverables
1	9/16	**Introduction:** What is Data Science? Doing Data Science Example: Formulating a Predictive Modeling Solution *Case Study: Data Science for Churn Reduction* ***Associated iPython Notebooks:*** Lecture_NumPyBasics, Lecture_PandasIntro, Lecture_SimpleiPythonExample, CaseStudy_Churn_Analysis	DSforBus Ch. 1,2 3 Elements Ch. 1 iPython Notebooks	**HW 1 Due 9/23**
2	9/30	**Data Analytic Thinking 1** Problem Formulation Data Sampling & Exploration Thinking through problem constraints *Case Studies: Exploratory Analysis & Visualization* ***Associated iPython Notebooks:***	DSforBus Ch. 4 Elements Ch. 2.1-2.2	
3	10/7	**Targeted Exploration:** **Guest Lecturer: Foster Provost** Exploring a Target Variable, Supervised Segmentation *Discussion: Target – Predicting Pregnancy* ***Associated iPython Notebooks:*** Lecture_DecisionTrees.ipynb, Lecture_Bagging_RandomForests.ipynb, Lecture_GradientTreeBoosting.ipynb	DSforBus Ch. 5 Elements Ch. 2.3 iPython Notebooks	**HW 2 Due**
4	10/14	**DS Hacking in Python** **Guest Lecturers:** Andreas Mueller, Robert Moakler	Elements 2.4, 2.6, 2.9, 4.4.1-4.4.2,4.5	**Project Team Assignments & Initial Ideas Due**
5	10/21	**Fitting a Mathematical Model to Data** Building, Understanding and Interpreting Models fit to a Loss Function *Case Study: Using Data Simulations* ***Associated iPython Notebooks:*** Lecture_ERM_LogReg.ipynb, Lecture_SVM.ipynb	Elements 3.1, 3.2, 4.4, 4.5 iPython Notebooks	**HW 3 Due**
6	10/28	**The Science of Predictive Modeling : Performance and Validation 1** How to design a predictive modeling experiment. Understanding why all models are wrong, though some are useful. *Case Study: Running a Modeling Bake-off and Proving Your Results*	DSforBus Ch. 7 Elements 7.1,7.2, 7.3 7.10	**Project Proposal Due**

		Associated iPython Notebooks: Lecture_SimpleOverfittingExample.ipynb Lecture_Resampling.ipynb, Lecture_Regularization.ipynb		

Class Number	Date	Topics	Readings	Deliverables
7	11/4	**The Science of Predictive Modeling : Performance and Validation 2** Selecting and understanding evaluation metrics, The Expected Value Framework, Cost sensitive metrics Learning Curves *Discussion: When you can't fit a model to customer satisfaction* ***Associated iPython Notebooks:*** Lecture_Metrics_Ranking.ipynb	DSforBus Ch. 8 iPython Notebooks	
8	11/11	**Advanced Data Preparation Topics** Data Prep for Predictive Modeling Feature Selection, Dimensionality Reduction Sampling Techniques *Case Study: Learning credit models under Concept Drift* ***Associated iPython Notebooks:*** Lecture_PhotoSVD.ipynb, Lecture_Binning_NonLinear.ipynb, Lecture_FeatureSelection.ipynb	DSforBus Ch. 11 Elements 3.3, 3.4 iPython Notebooks	**HW 4 Due**
9	11/18	**Similarity, Distance and Neighborhoods** Distances, k-Nearest Neighbors Collaborative Filtering Clustering ***Associated iPython Notebooks:*** Lecture_kNN.ipynb, Lecture_Clustering.ipynb	DSforBus Ch. 6	**Project Update Due**
	11/25	No Class (Thanksgiving Break)		
10	12/2	**Understanding Causality** *Case Study: Thinking through data products at LinkedIn*	DsforBus Ch 13, 14	**HW 5 Due**
11	12/9	**Big Data Science through the eyes of a Spam Filter:** Text Based Feature Engineering Naïve Bayes	DsforBus Ch 9, 10	
12	12/16	**Towards Analytical Engineering** Breaking problems into sub-problems. Learning through case studies. *Case studies: Etsy, Ebay, Dstillery*		**Project report due 12/19**
Final Quiz: *Taken online* ***by 11:59am on 12/22/13***				